Surveying the depth of user behavior profiling in mobile networks

Dominik Spörle, Marton Kajo*
*Chair of Network Architectures and Services, Department of Informatics
Technical University of Munich, Germany
Email: d.spoerle@tum.de, kajo@net.in.tum.de

***Abstract*—CDR (Call Data Records) allow network operators to collect location information, as well as other application level data about users of mobile devices. With various machine learning techniques it is possible to extract information to deduct general, or also very personal insights into the user's behavior. This survey presents various research topics working with CDR data, such as the evaluation of the mobility or social interactions of the users. This information is processed by using data mining methods, which yield patterns used to get insights in the user behavior.**

***Index Terms*—call data records, data mining, machine learning, mobility patterns**

1. Introduction

Mobile phones are popular devices for decades and used by a lot of most people daily. According to [1] in 2019 the number of mobile phone users is forecast to reach 4.68 billion. This entails a big amount of data. A CDR is a data structure created by network providers and stores relevant information about the telephonic activity. A CDR contains usually a spatial and a temporal resolution, which enables researches on the behavior of users by analyzing CDR data.
CDR data is often compared to GPS data, because it can track users in time and geographical space. It is important to note that GPS data is more precise and delivers more information about the movement of users, but CDR data is available since longer time and in a higher amount. Furthermore, CDR data can also contain information about the user's social interactions.
By analyzing patterns extracted from CDR data there are various areas to apply this knowledge. The understanding of mobility patterns yields insights into crowd analysis, population displacement, urban planning, network design, traffic management, targeted marketing, tourists movement or disease spreading. By analyzing the social interactions of calls between users it is possible to discover relationships, calling patterns or the social attributes of users. This information might help for example in areas like criminology to detect the social networks of criminals.
This paper presents and describes various techniques of creating patterns about the user behavior, while also highlighting the importance and usefulness of CDR data for many different areas by selected examples.
In Section 2 the processing of CDR data is described with different approaches, primarily showing methods to create movement trajectories of users by interpolating the spatial information of CDRs. In Section 3 different applications of CDR data is presented.

2. Handling & processing of CDR Data

To use CDR data to get insights in the behavior of users the data must be processed to get a movement trajectory of the users. Since CDRs only contain data of calls or texting activities of users, the accuracy of this data along the spatial and temporal dimension is limited, often referred to as spatiotemporally sparse data. To mitigate the sparsity of CDR data there exist several approaches of data completion, which try to fill the spatiotemporal gaps and derive movement trajectories from it. To do this the position of a user between the respective phone calls or sent text messages has to be estimated. This is done by applying movement models to the data with the use of machine learning techniques and data analysis. This is explained in the following sections.
Before the data of CDRs can be used for research, it is necessary to draw attention to privacy issues associated with CDR data. To work with it, the data must be anonymized to prevent that personal data can be inferred from it. Some of the approaches to do that are presented in Section 2.3.

2.1. Call Data Records

A call data record, or short CDR, contains in general the respective time of the call and the ID of the user. Also, it has an entry for the location, which is saved as the ID of the prefecture from where the call occurred. This information can be extended by appending data like the day of the week, the time of the previous call, etc. The structure of a CDR may depend on the communication service provider publishing the dataset.

2.2. Mobile Positioning based on CDR data

There are several approaches to refine CDR data and complete mobile positioning of users. As a first approach, the position of the user in between CDR events can be refined using a probabilistic model. In [2] an Inter-Call Mobility model (ICM) is introduced, which is based on a finite Gaussian mixture model.
The ICM model represents a spatiotemporal probability distribution of the location of a user between two consecutive CDR events. It relies on the Gaussian mixture model (GMM), which is a weighted sum of Gaussian Probability

doi: 10.2313/NET-2019-10-1_04

Density Functions (PDF). These functions represent the probability of finding a user at a position (x, y) at time t. The GMM is defined by the vector of all unknown parameters θ. Those parameters are estimated by the Expectation-Maximization algorithm [3], which performs a maximum-likelihood estimation. The number of components K in a Gaussian mixture is selected using criteria that combine the parameters estimation and a penalty that tries to prevent overfitting of the model. An initial estimate of the parameters θ is given by $K - means$ clustering.
In [2] the GMM is fitted to inter-call trajectories created during the data processing of the CDR data and the ICM model is defined. Applying the ICM model to two consecutive CDR events a probabilistic distribution of the user's position between the calls can be created. The the kernel density estimation of spatiotemporal probability distribution of user's inter-call movement is shown in Figure 1.

A second approach presented in [4], [5] is to create

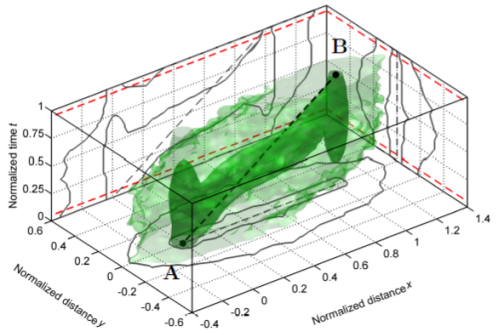

Figure 1: ICM model: probabilistic distribution of the user's position between the calls A and B, [2]

a model out of studies and analyzes of user's behavior. This is done by considering a user to stay more likely at certain locations, like his home or work place. The authors of this work separate the activities of users between a nighttime and a daytime period. To capture the locations for home and workplace the data is separated in those two periods and significant places are extracted. This is done by considering the place where the majority of CDR events during daytime occur to be the work place of the user and the place during nighttime to be his home.
For both day- and nighttime, different techniques are used for the completion of inter-call localization. During the nighttime period the user position is set to his home location if the last CDR is within some fixed temporal home boundaries. This method is expanded by adapting the nighttime interval and the home boundary for each user on previous observations made of him. For the completion of CDR data during daytime three factors are categorized, which affect the temporal cell boundaries of the user. Those factors are event-related, long-term behavior, and location-related. Event-related factors relate to the metadata of the CDR, like the duration of a call, what may give some indication if the user is static or moving. An example of a long-term behavior factor could be the number of unique visited locations, which can be related to the long-term mobility of the user. Location-related factors can yield to indications how relevant different places are in the user's activity.
With this information a model is created with supervised machine learning techniques to estimate the corresponding temporal cell boundary. The model is generalized from a training set consisting of CDR entries. Approximations are made with the Gradient Boosted Regression Trees technique [6].
In a third approach [7] the coverage area provided by the mobile operator and the location of the mobile device within this area is estimated. Also, the type of movement is detected in order to differentiate between a moving and staying user. Then, a map-matching technique is used to match the resulting location to a road (if a moving user is detected) or a building (if a stationary user is detected). To do this, first, the coverage area is estimated using the Voronoi method [8] and is optimized by comparing it with collected GPS data. This is done by minimizing a penalty function based on the observations of the GPS data. The coverage function is minimized with the implementation of the L-BFGS-B algorithm [9]. To differentiate between a moving and a staying user an implementation of a Kalman filter [10] is used and a mobility model is defined and integrated into this algorithm. For the map-matching of an estimated position, for each point corresponding road segments are found and matched. Therefore, a detection of road segments in a certain radius around the point is implemented and for each segment candidate points are computed with an orthogonal projection. The best candidate point is selected using the haversine distance [11]. A result of this map-matching step is shown in Figure 2. In general, it can be observed that the presented

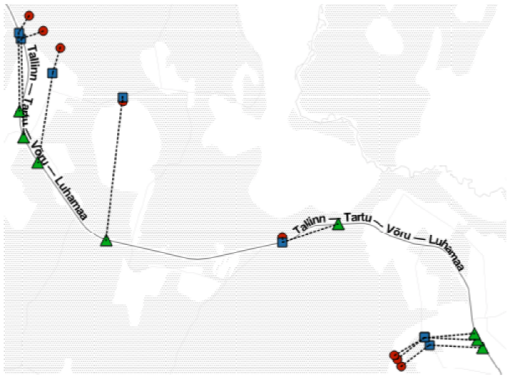

Figure 2: Map-Matching of CDRs: results of Kalman filter - red circles - estimated position, blue squares - map-matched position, green triangles - actual position, [7]

approaches yield to good results, if you consider that CDR data consist of sparse informations about the mobility of users. Estimated inter-call positions are not comparable to technologies like GPS, but they are precise enough to get useful insights into the behavior of users.

2.3. Anonymization of CDR data

As we can see in further chapters, CDR data can deliver personal information about users, which can cause privacy issues when working with it. It is challenging to find a good balance between the protection of the privacy of users and the utility of the data itself. Of course, published datasets do not contain any personal

information, but also anonymized data entails the possibility of re-identification of users. Traditional methods of anonymization are pseudo-anonymization (ID of the referenced user in the CDR is replaced with a code using cryptography), k-anonymity (trajectory of a user is hidden among $k-1$ other users with the same quasi-identifier) and spatial location cloaking (spatial noise is added to the data), but these methods are not efficient, because they either are not preventing re-identification or they impair the utility of the data.
An approach presented in [12] is to add time distortion instead of spatial distortion to the data. With a mechanism called Promesse the POIs (Points of interest) of users are hidden by smoothing the speed of the movements along the trajectories (see Figure 3). This approach is designed for mobility data in general, e.g. also GPS data, and is a good example of a method restricting the utility of data - especially if you apply it to CDR data. The protection of privacy is working, but one goal of CDR data processing is to find likely locations of users (POIs) and this approach is not very useful for this.
A more valuable anonymization method is described in

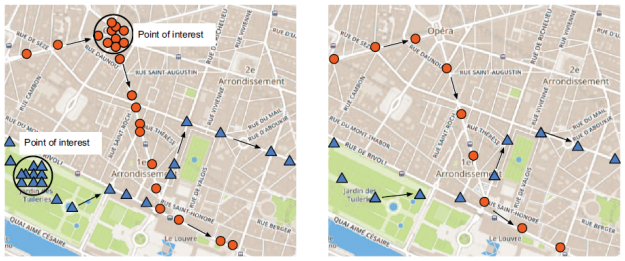

Figure 3: Overview of Promesse: smoothing the speed of movements - left: original dataset, right: after enforcing a constant speed, [12]

[13]. Here, the authors propose Differential Privacy (DP), a method that creates synthetic data out of the original one without any one-to-one correspondences between the two datasets. Their method DP-Star processes key factors and statistics learned from the original data and generates synthetic data out of it. This is done with 5 components of the system (which is shown in Figure 4). First an adaptive grid construction processes an effective discretization of the geographical location space of the dataset. With the trip distribution extraction, the correlation of the start and end point of a respective trajectory is kept. To mimic movements patterns of actual trajectories a mobility model is constructed, which is a Markov model. To estimate the route length of a synthetic trajectory a private median retrieval procedure is applied, which returns a noisy median of the trajectories. As last step the synthetic trajectory generation is processed in 5 separate steps: The start and end cells are generated by drawing a random sample from the trip distribution. The route length is determined by approximating it with an exponential distribution of the median lengths. The synthesizing of the trajectory as a sequence of cells is done with a random walk on the Markov mobility model. Finally, cells are converted to locations by randomly sampling a position with each cell. As a result a sequence of locations is the final trajectory.

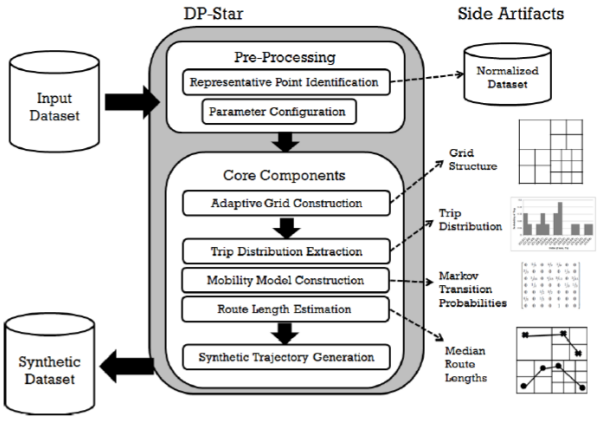

Figure 4: DP-Star architecture, [13]

A comparison of an original and a synthetic trajectory can be seen in the Figures 5 and 6.

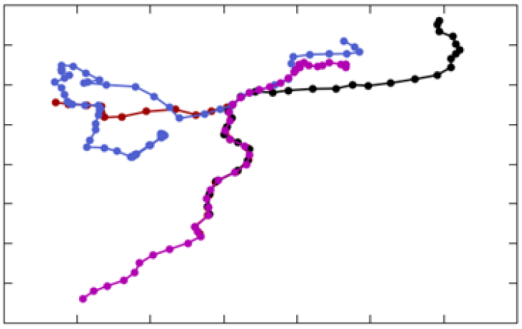

Figure 5: DP-Star: original CDR trajectory, [13]

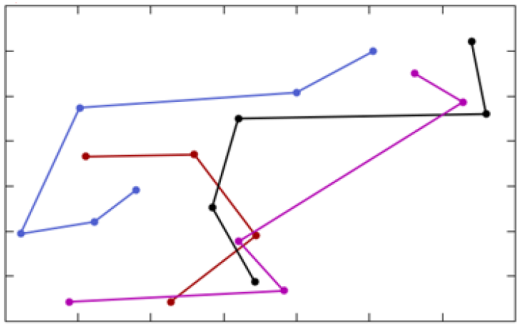

Figure 6: DP-Star: synthetic CDR trajectory, [13]

3. Insights

With the extraction of clusters from CDR data especially mobility patterns can give insights in the behavior of users. This can be used in a large variety of applications.

3.1. Data mining on CDRs

With the goal of characterizing users and extracting their behavior data mining analysis on CDR data is implemented. There are several different procedures for

doi: 10.2313/NET-2019-10-1_04

clustering and knowledge discovery. Often, they are based on analysis of the data itself, which yields to the definition of different subsets of features. For processing the CDR data can be structured as graphs, labeled sequences or sectioned vectors. The choice of the data structure can affect the results, it should be considered if for example topological information or computational resources are more important to the selected method.
A concrete data mining approach is presented in [14]. Here the cluster discovery is implemented with the LD-ABCD algorithm, which extracts separated clusters in the data and returns for each of the clusters the most appropriate local metric. This multi-agent algorithm is working on a weighted fully connected graph representing the data per agent. On each graph clusters are discovered by means of multiple Markovian random walks. These patterns are evaluated by a measure called Cluster Quality, which is dependent on the concept of graph conductance and the configuration of the dissimilarity measure. An agent might identify a set of similar clusters, which are merged to a meta-cluster. The output of the algorithm is a set of such meta-clusters. These clusters do not necessarily include the whole dataset. As a result, LD-ABCD is able to discover regularities and patterns among CDR data.

3.2. Analysis of tourism dynamics

In tourism data about the behavior is often created by analyzing interviews and surveys. In the area of big data social networks and economic datasets play a big role. Working with CDR data improves the quality of knowledge in various aspects of the tourism industry by extracting new indicators like, for example, the flow of tourists or profiles about different interests of tourists. These indicators can add value to the evaluation of touristic events and marketing strategies. In the approach presented in [15] indicators are mined from CDR data in the context of the country Andorra. As a result, the authors extract 6 different indicators by analyzing CDR data and comparing it to self-reported tourism surveys. The indicators found are:

- segment tourist flows (based on country of origin)
- new tourists and repeated tourists
- spatial distribution (based on country of origin)
- congestion
- revenues: gained by estimating the income of tourists in order to obtain the price of the mobile device used with the IMEI-TAC-code of the records
- tourists interest profiles: gained by comparing POIs of tourists with activities nearby respective cell towers

The authors then demonstrate in a case study how to add the value of those information to evaluations of tourism strategies by analyzing summer and winter events and tourist interests' profiles in Andorra. The results of those profiles can be seen in **Figure 7**.

3.3. Improved Quality of Experience with predictive models

For communication service providers (CSP) it is a big challenge to find patterns of customer behavior to improve

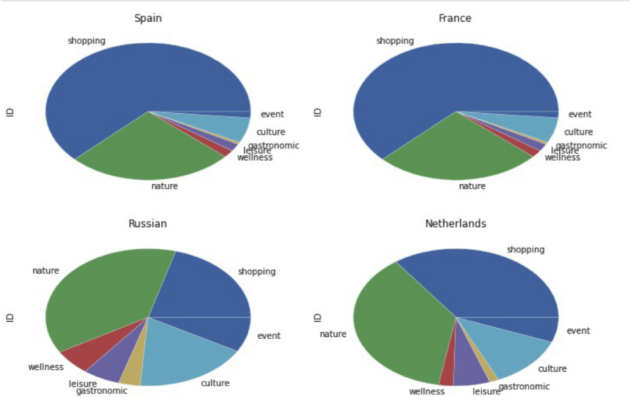

Figure 7: Tourism profiles: visualization of the nationality of tourists vs. their interests, [15]

the network and the customer's satisfaction. To assure a good level of Quality of Experience data management systems including Fault Management and Performance Management are required. To improve the performance of networks operators need analysis and diagnosis tools. Such a diagnosis tool is presented in [16]. This solution, which is called ARCD, Automatic Root Cause Diagnosis, is able to locate the root cause of network inefficiency. It uses logs which contain CDR data. Those CDRs are labeled either failed or successful and processed in several steps using equivalence class and graph computation. Finally an automated system diagnosing cellular networks based on the data collected from large-scale monitoring systems is implemented.
In [17] Customer Relationship Management (CRM) records are used to build a predictive model for customer churn (termination of the user's contract). To do that, CSP store customer transactions to discover patterns of customer behavior, which helps to find solutions to reduce their contract termination. CRM records contain contractual data. CDR data is required to complete it in order to predict churn. The approach is based on logistic regression [18] and random forest models [19].

3.4. Identifying criminal's behavior and social relations

Using CDR data it is possible to generate useful informations about the social relations and the behavior of users. Also, in the field of criminology these insights might be helpful to crack a criminal case. In [20] crime information and CDR data is combined to extract relationships and interactive patterns of criminal suspects.
This is done by implementing a knowledge graph analysis, which uses several graph traversal algorithms. The two datasets, CDR data and criminal cases, are both imported into a knowledge graph. The phone number of a single CDRs is implemented as nodes as well as the number of a criminal case. Edges are the call records itself, connecting phone nodes and the crime records connecting phones with related cases. The resulting knowledge graph is computed with a shortest path algorithm to discover shortest paths in the graph and thereby the contacting

doi: 10.2313/NET-2019-10-1_04

chains between users. Then the betweenness centrality algorithm is applied to the graph to extract the pivotal person in a social circle. With the Pagerank algorithm from Google (can be used as part of APOC library [21]) the importance of a node is calculated (as an example see Figure 8).
To detect different types of relationships a clustering model is created. To generate clusters features are defined based on the analysis of the respective data. In this approach features contain only the temporal component of CDRs, e.g. for example the duration of a call or the total number of calls during a specified time interval. With the features a Gaussian mixture model method computes the clustering model. As result 5 different clusters are extracted from the data, which contain unique characteristics. As an example, one of the resulting clusters has the properties that the contact of the users is kept for a longer period of time, they have a long holding time of their calls and the calls mostly occur in a working time and in relatively large numbers. This cluster reveals the relationship of the contacting users to be linked to their work. So, the model can extract the relationships and to separate these into possible categories like criminal confederates, working colleagues or close friends.

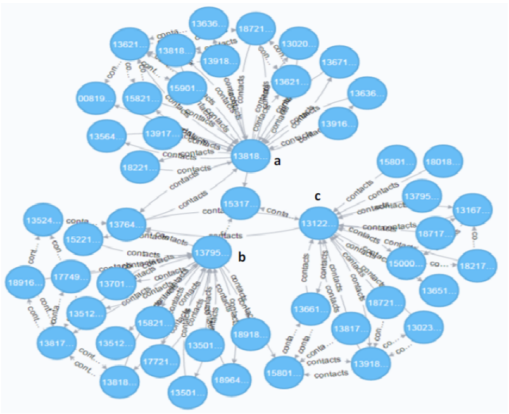

Figure 8: Page rank algorithm applied to a knowledge graph, [20]

4. Conclusion

It can be determined that CDR data can yield to useful information of the behavior of users. This can be used in a wide range of different topics. It can be concluded that especially the mobility patterns and the information about social interactions extracted by analysis of the data is valuable. Extracting information like the POIs of a user and his social relationships can be applied to improve research about human behavior and help to implement smart solutions for areas like marketing, urban transports, criminology, tourism and a lot more. To get good results working with CDR data it is often combined with other data, which can be geopositional or topic-related. Anonymization of CDR data is an important issue and current solutions do not face this issue in the way they should. On the other side approaches like DP-Star promise acceptable utility of the data and the protection of the privacy of users.

References

[1] Statista, "Number of mobile phone users worldwide from 2015 to 2020 (in billions) ," https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/, 2019, [Online; accessed 07-April-2019].

[2] M. Ficek and L. Kencl, "Inter-call mobility model: A spatio-temporal refinement of call data records using a gaussian mixture model," in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 469–477.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[4] S. Hoteit, G. Chen, A. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis," in *Proceedings of the Eleventh ACM Workshop on Challenged Networks*. ACM, 2016, pp. 45–50.

[5] G. Chen, S. Hoteit, A. C. Viana, M. Fiore, and C. Sarraute, "Enriching sparse mobility information in call detail records," *Computer Communications*, vol. 122, pp. 44–58, 2018.

[6] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[7] A. HADACHI, Amnir; LIND, "Exploring a new model for mobile positioning based on cdr data of the cellular networks." *arXiv preprint arXiv:1902.09399*, 2019.

[8] Wikipedia, "Voronoi diagram," https://en.wikipedia.org/wiki/Voronoi_diagram, 2019, [Online; accessed 12-May-2019].

[9] J. L. Morales and J. Nocedal, "Remark on" algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound constrained optimization"." *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 7–1, 2011.

[10] F. Xiao, M. Song, X. Guo, and F. Ge, "Adaptive kalman filtering for target tracking," in *2016 IEEE/OES China Ocean Acoustics (COA)*. IEEE, 2016, pp. 1–5.

[11] C. C. Robusto, "The cosine-haversine formula," *The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

[12] V. Primault, S. B. Mokhtar, C. Lauradoux, and L. Brunie, "Time distortion anonymization for the publication of mobility data with high utility," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1. IEEE, 2015, pp. 539–546.

[13] M. E. Gursoy, L. Liu, S. Truex, and L. Yu, "Differentially private and utility preserving publication of trajectory data," *IEEE Transactions on Mobile Computing*, 2018.

[14] F. M. Bianchi, A. Rizzi, A. Sadeghian, and C. Moiso, "Identifying user habits through data mining on call data records," *Engineering Applications of Artificial Intelligence*, vol. 54, pp. 49–61, 2016.

[15] Y. Leng, A. Noriega, P. A. S., I. Winder, N. Lutz, and L. Alonso, "Analysis of tourism dynamics and special events through mobile phone metadata." *arXiv preprint arXiv:1610.08342*, 2016.

[16] M. Mdini, G. Simon, A. Blanc, and J. Lecoeuvre, "Arcd: a solution for root cause diagnosis in mobile networks," in *2018 14th International Conference on Network and Service Management (CNSM)*. IEEE, 2018, pp. 280–284.

[17] K. A. NESTOR, Dahj Muwawa Jean; OGUDO, "Practical implementation of machine learning and predictive analytics in cellular network transactions in real time." *International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD) IEEE, 2018. S. 1-10.*, 2018.

[18] P. Bühlmann and B. Yu, "Boosting with the l 2 loss: regression and classification," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] Y. FAN, T. YANG, G. JIANG, L. ZHU, and R. PENG, "Identifying criminals' interactive behavior and social relations through data mining on call detail records," *DEStech Transactions on Computer Science and Engineering*, no. aiea, 2017.

[21] neo4j, "APOC Labrary," https://neo4j.com/developer/neo4j-apoc/, 2019, [Online; accessed 12-May-2019].